

RESEARCH STRATEGY

Background and significance

Protein glycosylation, the covalent attachment of a sugar moiety to an amino acid in the peptide chain, plays a key role in several important cellular processes, ultimately contributing to the physiology and

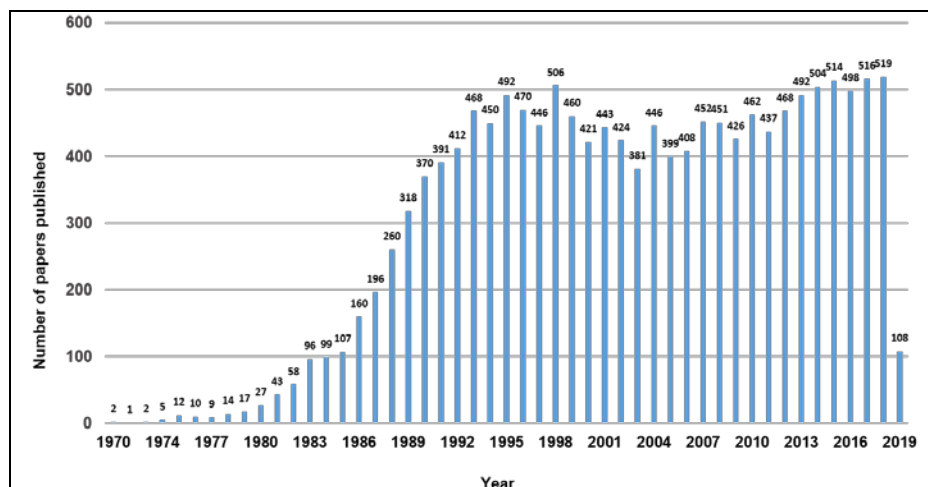


Fig 1a. The number of potential papers (14,949) with glycan information published over the past 40 years. The list of PubMed IDs was compiled using the query – [(glycosylation[Text Word]) AND site[Text Word]] (sorted by “best match”) in PubMed database (as of 3/27/19). This query provides publications where text words ‘glycosylation’ and ‘site’ are present together or separately.

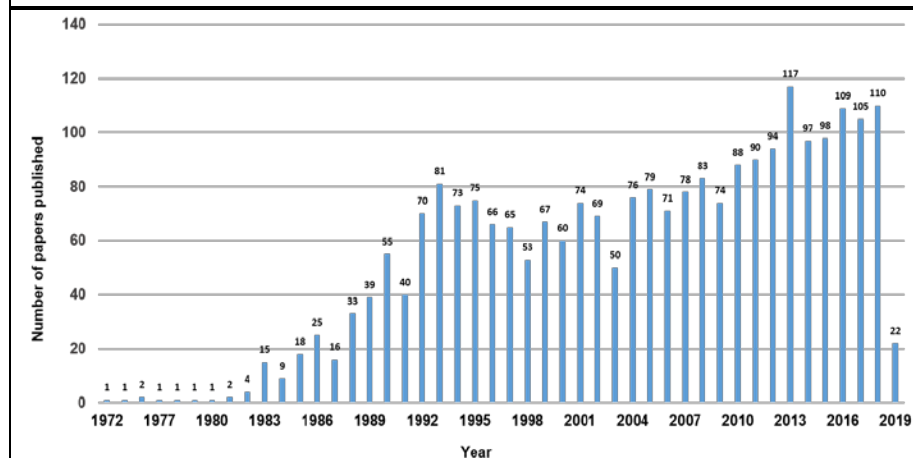


Fig 1b. The graph represents the potentially relevant papers (2,511) compiled using query- [("glycosylation site" OR "glycan structure" OR "glycan composition")] (sorted by best match) in PubMed database (as of 3/22/2019). The double quotes in the query ensures exact match (in the same order) of the text words.

important to capture different experimentally determined glycoforms reported in publications, as an effort to capture context specific components of glycosylation (such as biological source, protein sequence, variation, and glycan (micro) heterogeneity for each site) and to potentially highlight its role in diseases. Given the increased interest in understanding the role of glycosylation, the number of glycobiology specific publications have increased over the last several years (see Fig. 1a, 1b). Fig. 1a provides a

development of an organism [1]. One of the most common and important feature of protein glycosylation is microheterogeneity. It suggests that multiple glycan structures can be observed at a specific glycosylation site on the same protein expressed in the same cell type. This variation can increase dramatically when extended to multiple sites on the same protein or the same protein from different biological sources (e.g. different cell type). Thus, the same protein can exist in different, context dependent “glycoforms” [2]. This results in a distinct glycome for each microenvironment [2], making glycosylation one of the most complex post-translational modifications to study. Changes in glycosylation can contribute to different diseases such as cancer [3]. Recent studies such as the one led by Li et al. [4], which shows the possibility for targeting glycosylated PD-L1 with a drug-conjugated antibody is opening up highly promising areas of immuno-oncology research where glycans play a central role. Therefore, it is

generic search result, whereas Fig. 1b shows results using additional search conditions. The number of publications, with relevant information most likely lies somewhere in between the two results. Capturing and presenting site-specific protein-glycan data in addition to other context specific information from these publications can assist GlyGen users wishing to explore the complexity of this biological process in context of relevant genomic, pathway and other molecular biology evidence.

INNOVATION

Innovation includes development of a scalable new text-mining approach for translating experimental glycobiology data into glycobiology knowledge. The approach includes development of a workflow that encompasses automatic text-mining of glycosylation related data as well using different tools to provide novel annotations through inference. The developed workflow, including the text-mining tool, will be available for reuse as an effort to allow constant addition to knowledge by providing the necessary platform to capture all aspects of glycosylation reported through publications. The proposed project is expected to not only expand the knowledge available through GlyGen but also improve the quality of data, thereby increasing the usefulness of the GlyGen resource.

APPROACH

Team

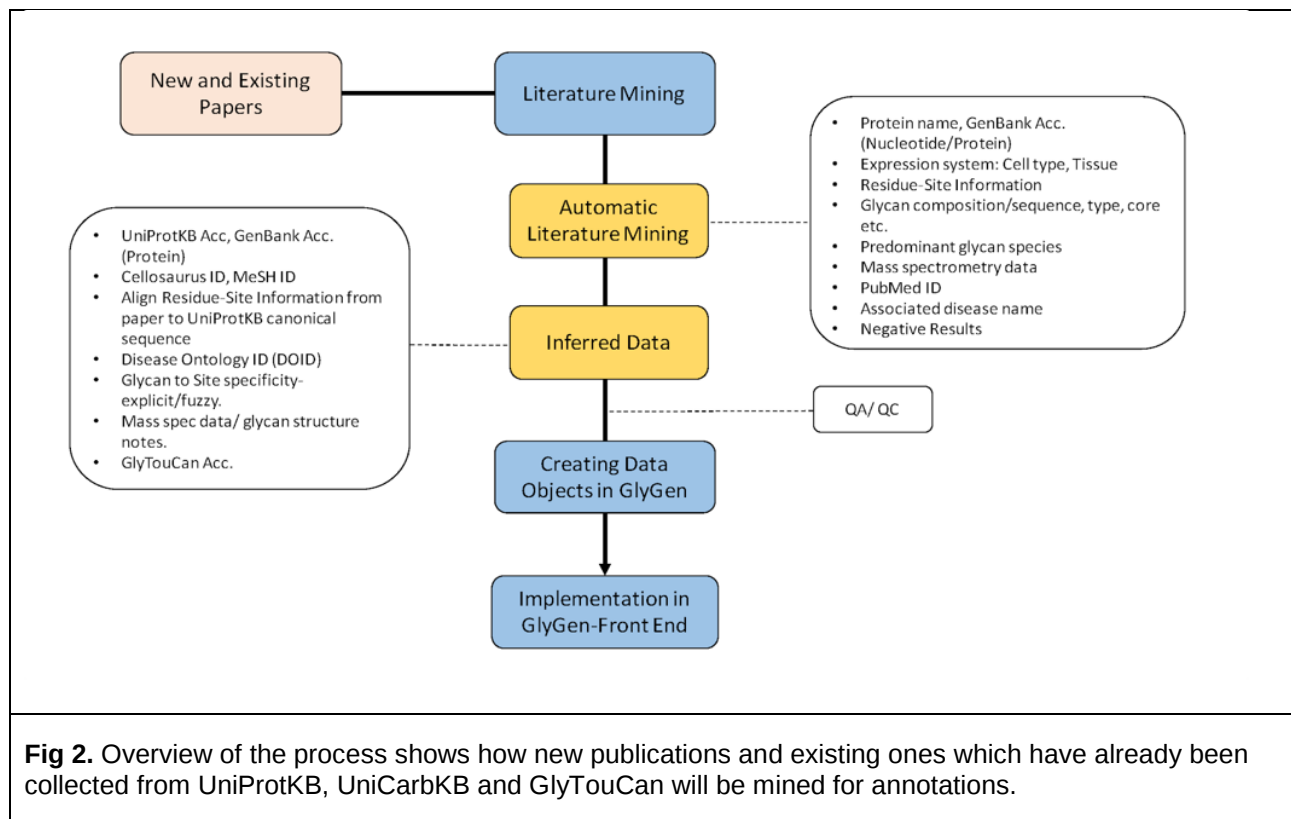
The team includes Dr. Vijay Shankar (University of Delaware) with extensive experience in developing text mining algorithms and Dr. Raja Mazumder (George Washington University) with expertise in data integration and management. Dr. Shankar and Dr. Mazumder have co-authored papers in literature mining and data integration. Dr. Shankar and Dr. Mazumder will work closely with the GlyGen team to ensure success of the project.

Aim 1: Develop and run text-mining software that can extract information from publications which can be used to improve GlyGen content.

Aim 1a) Develop and use automatic text-mining tool on the existing publications in GlyGen (sourced from UniProtKB, UniCarbKB, GlyTouCan) as a method to add evidence tags to existing annotations as well as capture additional information.

The tool will capture specific data from the publications (title, abstract, and PubMed Central (PMC) open access full-length publications). This data will be compared with existing annotations in GlyGen and whenever possible evidence tags will be added to existing information from UniProtKB, UniCarbKB and GlyTouCan. Annotations such as expression type (tissue, cell type etc) of the glycoprotein, proteolytic enzymes, connection to disease, glycan (composition, structure), mass spectrometry data (such as analyzer, sample preparation method, mass of glycan and glycoprotein etc.) through tables and figures will also be captured wherever available in publications. Additional data such as UniProtKB accession, human disease ontology ID will be automatically added through inference based on algorithms developed earlier by our group [5]. The overview of the entire process is shown in Fig 2.

We aim to retrieve the selected information from 801 papers in GlyGen (as present on 12/5/18 from UniProtKB, UniCarbKB and GlyTouCan). Literature mining on these existing GlyGen papers will add new annotations such as expression system of the glycoprotein, glycan heterogeneity at specific sites, predominant glycan species, negative information (for e.g. absence of glycosylation) and



correspondence of these factors to disease, most of which has not been previously captured in GlyGen. In addition to these new annotations, literature mining will also provide additional level of evidence that can be associated with existing annotations, generated by several curators or automated sources over different periods of time in UniProtKB, UniCarbKB and GlyTouCan. Comparison of the results provided by the text-mining tool with annotations already captured in GlyGen will allow us to identify discrepancies and to determine whether each is programmatic (due to tool malfunction) or simply incorrect in the data sources we use to populate GlyGen.

We plan to extract the following information with respect to glycosylation for human [taxid:9606] and mouse [taxid:10090] species and over time extend them to other organisms:

Data captured through Automatic Literature Mining:

1. Glycosylated protein
 - a. Protein name
 - b. Cell-type or tissue in which the protein was expressed (Expression system)
2. Glycosylation site for the protein
 - a. Residue (e.g. Ser or Thr or Asn)
 - b. Position
 - c. Type of linkage/ glycosylation (e.g. N-linked or O-linked).
3. Glycan information
 - a. Composition (format provided in the paper. For eg. HexNAc2Hex3)

- b. Glycan core type (eg. Hybrid / Complex / High-Mannose)
 - c. Link sugar (e.g. GlcNAc, GalNAc)
 - d. Predominant/ most abundant glycan species. (with eventual mapping to GlyTouCan/PubChem/ChEBI accessions).
4. Mass spectrometry analysis information/details: (Glycopeptide/glycan)
 - a. Sample preparation method
 - b. Mass
 - c. Analyzer
 - d. Ionization method
 - e. Chromatography matrices and protocols
 - f. Lectin characterization
5. Evidence:
 - a. PubMed ID
6. Connection to disease (e.g. type of cancer).
7. Negative results (e.g. Potential glycosylation site which was never found to be glycosylated).

Data inferred through new and existing tools developed by us:

1. UniProtKB [6] accession or RefSeq [7] Accession
2. Mapping residue position from paper to UniProtKB canonical sequence
3. Cellosaurus ID [8]
4. Human Disease Ontology ID (DOID) [9]
5. GlyTouCan Accession [10] or PubChem [11] or ChEBI [12].
6. Glycan sequence (e.g. IUPAC or GlycoCT [13] or WURCS [14])
7. Glycan-site specificity ('site-specific' if the a single glycan is always attached to a specific site on a protein or 'multiple' if more than one glycan is present on a particular glycosylation site or multiple sites bear the same glycan)

Development of the text mining tool for this project will be based on our experience in developing RLIMS-P (Rule-based Literature Mining System for protein Phosphorylation), currently being extended

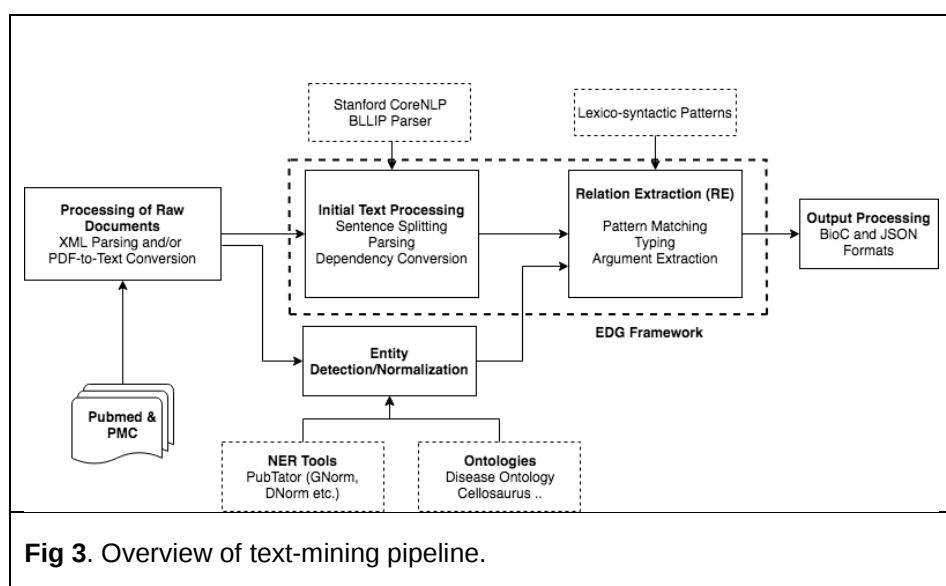
Table 1: Performance and functionality of our published relation extraction tools.

System	Precision	Recall	F-measure
RLIMS-P	0.88-0.96	0.88-0.94	0.88-0.95
eFIP	0.92	0.77	0.84
miRTex	0.94-0.96	0.81-0.91	0.88-0.94
eGARD	0.93-0.95	0.74-0.86	0.90-0.97
miRiaD	0.95	0.84-0.90	0.89
DiMeX	0.87-0.95	0.85-0.89	0.88-0.91

for other PTMs. To extract information from research literature on a large-scale, we have developed several relation extraction (RE) tools. Some of these tools relevant to this project have been applied on entire Medline and entire PMC open access (OA) articles to extract specific relations. We have successfully developed several PTM-focused text mining tools (see Table 1) using our state-of-the-art Natural Language Processing (NLP) approaches, including **RLIMS-P** [15-18] for mining phosphorylation events (kinase, substrate, and site) and **eFIP** [18-20] for mining the impact of phosphorylation on substrate protein interactions. The tools have achieved high precision, have been evaluated in the BioCreative Interactive Task for usability and utility [17,20,21] and are being adopted for computer-assisted literature-based

curation by a growing number of databases and ontologies (e.g., Phospho.ELM [22], PhosphoGRID [23], Gallus Reactome [24] and PRO [25]). To capture disease/cancer-related context, we have developed two extraction tools: **miRiaD** [26], which extracts miRNA target and expression [27] as well as the impact on processes in the disease context, and **DiMeX** [5], which mines disease associations with mutations. Additionally, **DEXTER** [28] was developed to extract differential expression in disease/tissues and **eGARD** [29] to extract association of genomic anomalies with drug responses.

Based on our extensive experience in developing algorithms, tools, and resources for relation extraction (RE) and mining information from text, we have created a general framework named Extended Dependency Graph (EDG) [30] that incorporates linguistic principles to move from syntactic dependencies to more semantic predicate-argument dependencies that are crucial for various RE systems. EDG framework allows fast development of rule-based systems and helps machine learning-based RE systems as it abstracts away from textual variations when it builds its representation. Based on the workflow (see Fig. 3) we will use the documents in the form of Medline abstracts or full-length PMC articles which will be converted to text using pdf to text conversion or XML parsing (if the text is a PMC open access article). These texts will undergo initial processing that includes splitting the text into sentences and tokenization. Additionally, the relevant text will be parsed using the Charniak–Johnson parser [31,32], with David McClosky's adaptation to the biomedical domain [33] .



The parsed structure will be converted into dependencies that will be manipulated in the RE stage. In addition, certain extra-syntactic information will be added to dependencies that will be useful for relation extraction. Simultaneously, using different named-entity recognition (NER) tools and normalization tools, various type of entity mentions (including genes, proteins,

diseases, cell lines, tissues, modification sites) will be detected and normalized to standard ontological IDs. The text with dependencies and named entities will be used as input for our relation extraction framework, which we developed in-house using EDG. This RE framework requires relation specific lexico-syntactic patterns defined on lexical triggers, syntactic dependencies and the types of the arguments. Additional processing may be required to extract the exact argument from matched positions (e.g. disease from a named cell-line or from a genetic disease mention). The output can be produced in standard formats such as BioC [34] and JSON (<https://www.json.org/>).

In most of our tools, we have used a combination of a few methods (in-house, **GNorm/Pubtator** [35]) for gene and protein name detection and normalization. Protein site detection will be done using the technology we have developed for RLIMS-P. We have been extending RLIMS-P using the EDG relation extraction framework to other post-translational modifications, including glycosylation. This work will be

completed, comprehensively evaluated and used for this project. We will need to extend this system to capture the type of linkage information. While developing the adaptation of RLIMS-P technology to extract glycosylation information, we have conducted preliminary work on glycan information such as the sugar and recognition of glycan chemical structure.

DEXTER's relation extraction module will be applied with minor modifications to tailor it to the extraction of the cell-type/tissue expressing the glycosylated protein. The connections to disease (e.g. cancer) will be extracted based on techniques that we had developed in the context of eGARD and miRiaD, which also include algorithms for disease selection and normalization to DOIDs. We have developed a modular negation detection system enabling it to be plugged into any EDG-based RE system with minimal effort. This negation system will be used to extract negative glycosylation information. The results provided by the tool will be compared to the existing annotations in GlyGen which will be classified as annotation tool errors or errors in GlyGen. The process to extract glycan information in a comprehensive manner will be a collaborative effort between researchers at George Washington University and University of Delaware and the rest of the GlyGen team.

Aim 1b) Increase site-specific and protein specific glycan data in GlyGen by adding new annotations, glycoproteins, glycosylation sites, and glycan structures through literature mining of new publications (publications not yet in GlyGen).

The tool will be developed, refined and extended to 3000 new publications (a conservative approximation based on results shown in Fig. 1a and 1b and additional publications available from previous research in text-mining related to glycobiology [36-38]). The data captured through the automatic literature mining and through inference will be presented as GlyGen data objects after integration QA/QC process. Table 2 provides examples of the anticipated (extracted and inferred) data from publication.

Table 2. Example row from publication (PMID: 8142896) [39]. The underscored text represents automatically inferred data using tools described above. (This table is prepared manually for the purpose of this proposal).

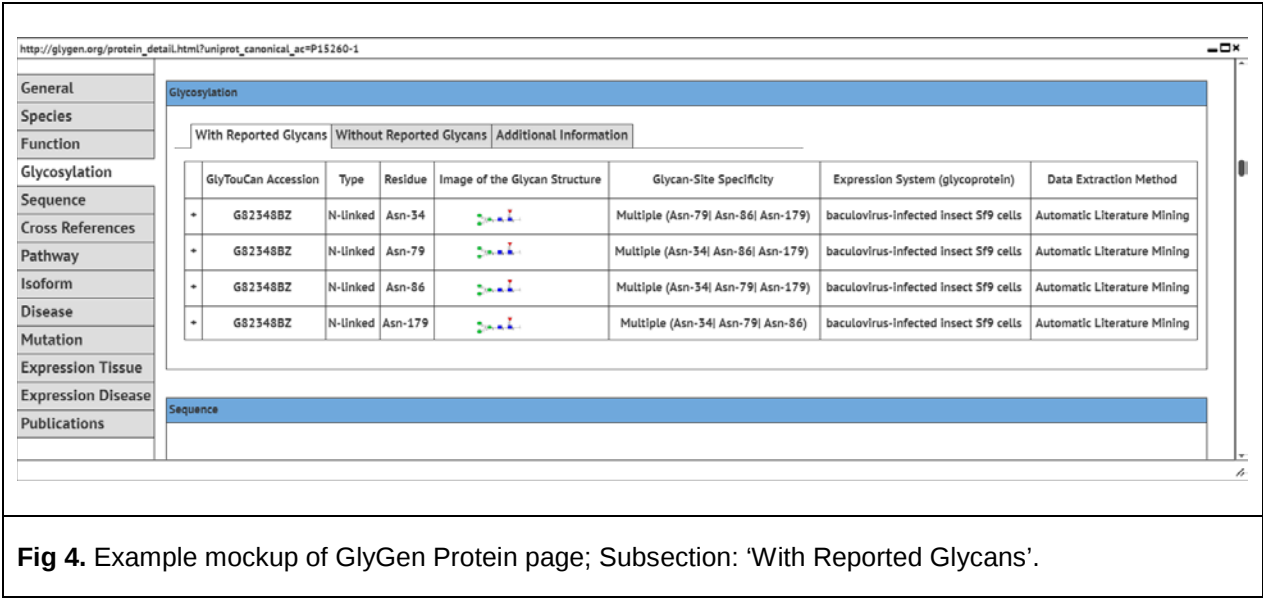
Field	Value
Glycosylated protein-	
Protein name	human interferon γ receptor
UniProtKB accession	<u>P15260-1</u>
GenBank accession	
Mass	
Expression system (protein):	
Tissue	
Cell line	baculovirus-infected insect Sf9 cells
Cellosaurus ID	-
Glycosylation site-	
Amino acid (Residue)	Asn
Position (position in the paper)	17

Position (position in UniProtKB canonical sequence)	<u>34</u>
Type of linkage/glycosylation	N-linked
Glycan-	
Glycan sequence	Man α 1,6(Man α 1,3)Man β 1,4GlcNAc β 1,4(Fuc α 1,6)GlcNAc
Glycan composition (if glycan sequence not fully defined)	-
Glycan core (if no glycan sequence)	-
Link sugar (if no glycan sequence)	-
Predominant/ abundant glycan species.	The predominant species was a hexasaccharide of molecular mass 1,039, containing a fucose subunit linked to the proximal N-acetylglucosamine residue.
GlyTouCan/PubChem/ChEBI accession	-
Mass spectrometry analysis information/details: <i>(will be presented as separate columns specific to glycan/glycopeptide)</i>	
Mass (glycan)	1039
Sample preparation	endoproteinase Lys-C endoproteinase Glu-C proteinase K
Glycosidase treatments	N-glycosidase F
Analyzers	ion spray mass spectrometry on an API 111 system (SCIEX) matrix-assisted laser desorption ionization mass spectrometry on an LDI-1700 time-of-flight mass spectrometer
Ionization method	-
Chromatography matrices and protocols	high-performance anion-exchange chromatography
Lectin characterization	-
Glycan-site specificity (position in the paper)	Multiple (Asn-62 Asn-69 Asn-162)
Glycan-site specificity (position in UniProtKB canonical sequence)	<u>Multiple (Asn-79 Asn-86 Asn-179)</u>
Connection to disease	
DOID	
Evidence (PMID)	8142896
Negative results (position in the paper)	Asn 223 was never found to be glycosylated
Negative results (position in UniProtKB canonical sequence)	Asn <u>240</u> was never found to be glycosylated

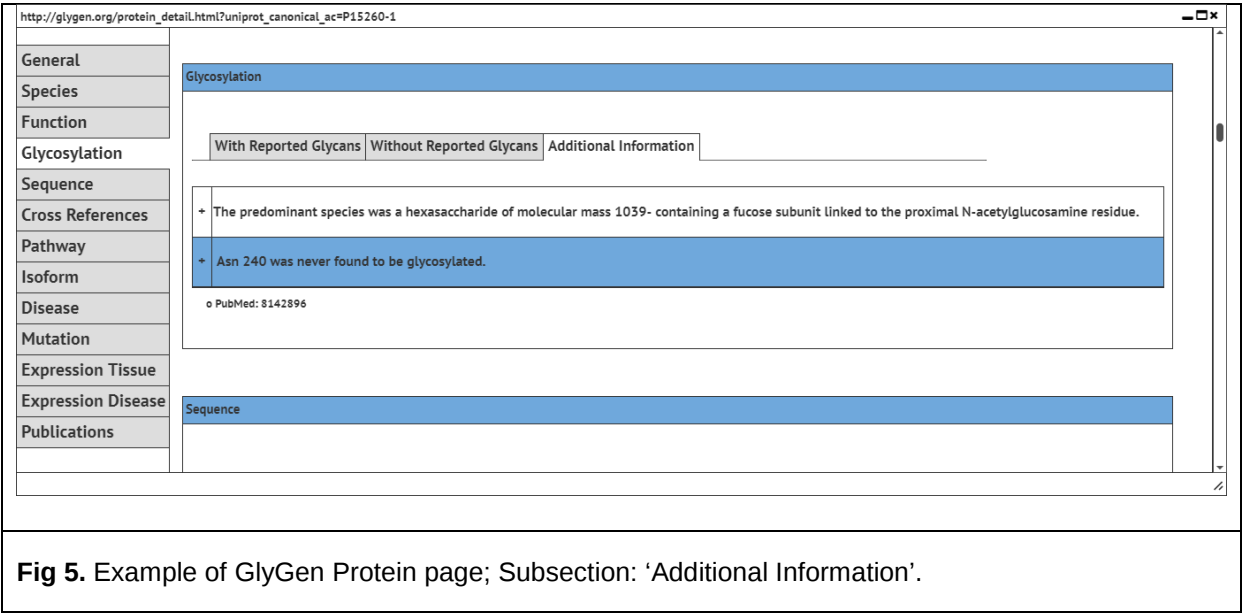
Presentation of the results in GlyGen

New annotations will be integrated into GlyGen through new or existing sections/subsections on GlyGen Protein and Glycan pages. Examples for some of the data representation in GlyGen are shown below.

Glycan-site specificity and expression system (glycoprotein) will be integrated into the existing subsection (With Reported Glycans) of the “Glycosylation” section and will be visible through GlyGen Protein page (Fig 4).



Information such as “Predominant glycan species” and “negative information” will be represented as literature comments through a new subsection (Additional Information) under “Glycosylation” section (Fig 5).



Leveraging our experience in developing tools such as DiMex [5] and DEXTER [28] we plan to capture disease information from the papers which will potentially help provide information related to causation or co-relation of a glycan or glycoprotein to the associated disease. The disease terms will be further mapped to the Human Disease Ontology ID’s. With this approach we intend to capture the involvement

of the glycoprotein in a particular disease by capturing the change in glycosylation state (for e.g. site, glycan structure or composition) in diseased cells/tissues. This data along with mass spectrometry data, will be represented in the existing protein page.

All of the above information will be searchable and captured in BioCompute Objects format [40] as evidence. Final presentation of literature mined data will be decided based on user and GlyGen team member input.

Expected outcomes, potential pitfall and alternative strategies

Based on our experience in developing text-mining algorithms as well as data integration, we intend to provide an extensive glycan-protein data and annotations from existing publications in GlyGen and also publications currently not yet present in our resource. All information will have detailed provenance information captured in BioCompute Objects [40] and available through data.glygen.org. Because of the complexity and heterogeneity of glycosylation and the way they are described by different authors, the extraction of the data from publications without losing information is a key challenge for this project. To overcome this, reliability metrics will be provided that will help users filter out potential false positive results. Metric rules will be determined manually through spot-checking, but they will be applied automatically, and hence will be scalable.

Aim 2: Replicability and usability of the tool.

Aim 2a) Generate a list of glycan specific terms from publications that can be subsequently leveraged to identify the structural features or compositions of the reported glycans.

Our goal is to identify and extract mentions in the text that are suggestive of the glycan composition or structure. From these mentions, we will generate a list of glycan strings that can form the basis for ultimately automatically mapping the glycan structure (in IUPAC or any other similar sequence format) to GlyTouCan Accession (future plan). This will provide a reference allowing glycan composition to be mapped to GlyTouCan Accession, which would facilitate deeper inference.

For the purpose of extracting mentions of glycan composition or structure from the text, we aim to employ pattern-based (regular expression) methods augmented with contextual clues that indicate the presence of glycan structure reference. The extraction method will be refined in an iterative manner with feedback from GW researchers. We intend to utilize GlyGen hackathon/workshop to assign GlyTouCan/PubChem/ChEBI accessions manually during the project period. This list will be the initial step in creating a dictionary to assist GlyGen users and improve representations in chemical databases. Table 3 provides an example draft list. The final format of the list will be determined based on input from users and GlyGen team.

Table 3. Examples of captured sentences used to describe glycan(s) reported in the papers (identified by PMID). The glycan structural or composition features are underscored. This list will serve as the initial step in generating a potential dictionary. *Note: See Aim 1 for additional contextual details that will be extracted from publications.*

PMID	Source text from publications
19088065	Our main finding is that KLK6 from ovarian cancer ascites (but not CSF) is <u>extensively sialylated</u> .

	These results suggested presence of <u>terminal sialic acid</u> residues on ascites-derived KLK6 but not on the CSF-derived KLK6.
	The majority of the identified structures were <u>core-fucosylated bi-, tri-, or tetra-antennary glycans</u> with a varying number of <u>terminal galactose-linked sialic acids</u> .
19343721	The N-glycans of hPC are <u>complex di- and tri-sialylated structures</u> <u>monosialylated biantennary glycan</u> with a <u>core fucosylation</u> .
2813359	However, structural analyses of their N-linked sugar chains revealed that EPO-bi contains the <u>bi-antennary complex</u> type as the major sugar chain, while EPO-tetra and the standard EPO contain the <u>tetra-antennary complex</u> type as the major sugar chain.
24706782	Detailed glycosylation analysis exhibited <u>complex and oligomannosidic N-glycans</u> in a site-specific manner on human-serum IgM and on plant- and human-cell-line-produced SM6.

Aim 2b) Provide literature mining software as a deliverable so that it can be run automatically on new publications.

Docker Container along with appropriate documentation will be created as an effort to package this tool for easy usage. Major tasks include:

1. Establishing a Docker container in GlyGen Github repository as a way to provide the ability to use the tool on new publications.
2. Provide supporting documentation.

Almost all our text mining tools such as RLIMS-P, eFIP, and eGARD have been dockerized, with these versions being checked out, downloaded and deployed at different locations. Leveraging this experience, we will provide docker containers for the literature mining software developed that will enable portability and parallel execution of tools on massive datasets (e.g. in AWS), yielding outputs faster. Additionally, this docker will provide the ability to perform periodical automatic updates of text mining results for new publications. The docker and the related source code for the tool will be made available in GlyGen GitHub with supporting documentation and usage instructions for easy use and deployability.

Expected outcomes, potential pitfall and alternative strategies

Literature mining software with documentation will be provided. Additionally, a list of terms used in publications to describe glycans will be automatically created. It is possible that the terms extracted automatically from publications would require evaluation. Therefore, these terms will be reviewed in GlyGen hackathon/workshop/face-to-face meetings.

Execution, implementation and milestones

The Gantt chart (Fig. 6) provides the timelines for the key tasks and deliverables that we plan to achieve during this project. The initial phase will be focused on developing and testing the text-mining tool to capture data from the existing papers in GlyGen and to provide its availability through GlyGen interface. The later phase will focus on extending this tool to new papers, parallel with development and refinement of the process. Other key tasks include validation and quality control of the data, addition of new

annotations through inference as well as setting up the necessary structure to provide reusability of the text-mining tool. Dr. Mazumder will be responsible for providing examples of true/false positive/negative

results and iterative testing of the results and tool. Dr. Shanker will be responsible for tool development, running of the tool and providing the tool at the end of the project period.

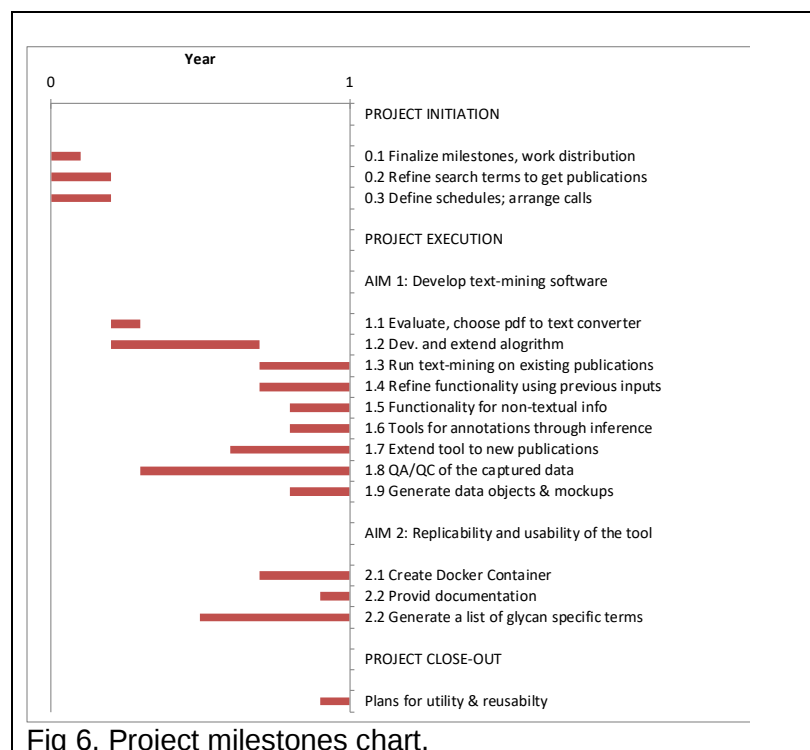


Fig 6. Project milestones chart.